

b)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2000-101442
(P2000-101442A)

(43) 公開日 平成12年4月7日(2000.4.7)

(51) Int.Cl. ⁷	識別記号	F I	テマコード*(参考)
H 0 3 M 7/40		H 0 3 M 7/40	
G 0 6 F 5/00		C 0 6 F 5/00	H

審査請求 未請求 請求項の数26 O L (全 31 頁)

(21) 出願番号	特願平10-272724	(71) 出願人	000005223 富士通株式会社 神奈川県川崎市中原区上小田中4丁目1番1号
(22) 出願日	平成10年9月28日(1998.9.28)	(72) 発明者	森原 隆 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
		(72) 発明者	矢作 裕紀 神奈川県川崎市中原区上小田中4丁目1番1号 富士通株式会社内
		(74) 代理人	100079359 弁理士 竹内 進 (外1名)

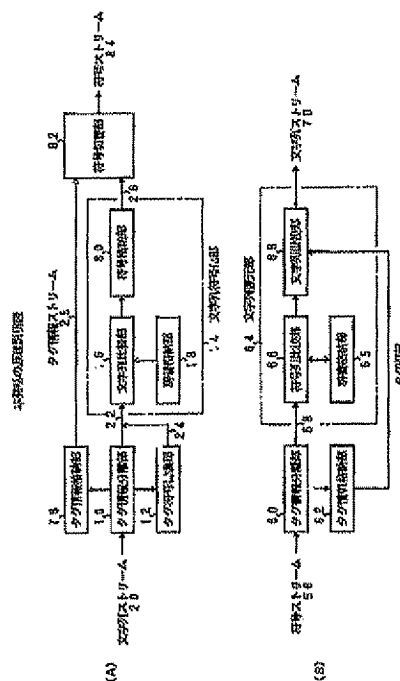
最終頁に続く

(54) 【発明の名称】 データ圧縮装置及び復元装置並びにその方法

(57) 【要約】

【課題】 タグを含む構造化文書を、文書の検索あるいは読み込み時間の短縮とメモリやディスク容量の増加を最小限とするように圧縮し復元する。

【解決手段】 タグ情報分離部10は、文字列ストリームから識別したタグを分離してタグ情報として出力する。タグ置換部12は、タグが分離された文字列ストリームの位置に識別のためにタグ符号を配置する。文字列符号化部14は、タグ符号置換部12から出力されるタグ符号を含む文字列ストリームを圧縮符号化する。データ復元装置は、圧縮と逆の操作で文字列ストリームを復元する。



【特許請求の範囲】

【請求項1】 タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部と、
前記タグ情報分離部でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換部と、
前記タグ符号置換部から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化部と、を有することを特徴とするデータ圧縮装置。

【請求項2】 請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮装置。

【請求項3】 請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、前記タグ情報分離部で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮装置。

【請求項4】 請求項1記載のデータ圧縮装置に於いて、更に、
前記タグ情報分離部で分離されたタグ情報を格納するタグ情報格納部と、
前記文字列符号化部で生成された符号データを格納する符号格納部と、
前記タグ情報格納部に格納されたタグ情報と符号格納部に格納された符号データを選択して出力する符号切替部と、を設けたことを特徴とするデータ圧縮装置。

【請求項5】 請求項1記載のデータ圧縮装置に於いて、前記文字列符号化部は、
圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部と、
前記タグ符号置換部からの文字列ストリームの中の部分文字列と前記辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部と、を備えたことを特徴とするデータ圧縮装置。

【請求項6】 請求項1記載のデータ圧縮装置に於いて、更に、前記タグ情報分離部で分離したタグ情報を圧縮するタグ情報圧縮部を設けたことを特徴とするデータ圧縮装置。

【請求項7】 請求項1記載のデータ圧縮装置に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部で分離したタグ情報に含まれる文字

列ストリームの部分文字列と前記タグ辞書格納部の登録文字との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部と、を備えたことを特徴とするデータ圧縮装置。

【請求項8】 請求項4記載のデータ圧縮装置に於いて、更に、前記文字列符号化部で生成した符号データの中のタグ位置を検出するタグ位置検出部を設け、前記タグ情報格納部に前記タグ情報分離部で分離したタグ情報と共に前記タグ位置検出部で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮装置。

【請求項9】 請求項8記載のデータ圧縮装置に於いて、前記タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納部にタグ情報と共に格納したことを特徴とするデータ圧縮装置。

【請求項10】 タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置に於いて、

前記符号ストリームからタグ情報と符号データとを分離するタグ情報分離部と、
前記タグ情報分離部で分離したタグ情報を格納するタグ情報格納部と、
前記符号データから文字列及びタグ符号を含む文字列データを復元した後に、前記タグ符号をタグ情報格納部のタグ情報に置き換える文字列復元部と、を備えたことを特徴とするデータ復元装置。

【請求項11】 請求項10記載のデータ復元装置に於いて、前記文字列復元部は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書格納部の参照で元の文字列を復元する文字列比較部と、

前記文字列比較部により復元したタグ符号を、前記タグ情報格納部のタグ情報に置き換える文字列置換部と、を備えたことを特徴とするデータ復元装置。

【請求項12】 請求項10記載のデータ復元装置に於いて、更に、前記タグ情報格納部に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設けたことを特徴とするデータ復元装置。

【請求項13】 請求項10記載のデータ復元装置に於いて、更に、

復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部と、を備えたことを特徴とするデータ復元装置。

【請求項14】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、
前記タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、
前記タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を有することを特徴とするデータ圧縮方法。

【請求項15】請求項14記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮方法。

【請求項16】請求項14記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、前記タグ情報分離過程で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮方法。

【請求項17】請求項14記載のデータ圧縮方法に於いて、更に、
前記タグ情報分離過程で分離されたタグ情報を格納するタグ情報格納過程と、
前記文字列符号化過程で生成された符号データを格納する符号格納過程と、
前記タグ情報格納過程に格納されたタグ情報と符号格納過程に格納された符号データを選択して出力する符号切替過程と、を設けたことを特徴とするデータ圧縮方法。

【請求項18】請求項14記載のデータ圧縮方法に於いて、前記文字列符号化過程は、
圧縮する際の処理単位となる文字列を登録した辞書を生成する辞書生成過程と、
前記タグ符号置換過程で得られた文字列ストリームの中の部分文字列と前記辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。

【請求項19】請求項14記載のデータ圧縮方法に於いて、更に、前記タグ情報分離過程で分離したタグ情報を圧縮するタグ情報圧縮過程を設けたことを特徴とするデータ圧縮方法。

【請求項20】請求項14記載のデータ圧縮方法に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書の登録文字

列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。

【請求項21】請求項17記載のデータ圧縮方法に於いて、更に、前記文字列符号化過程で生成した符号データのタグ位置を検出するタグ位置検出過程を設け、前記タグ情報分離過程で分離したタグ情報と共に前記タグ位置検出過程で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮方法。

【請求項22】請求項21記載のデータ圧縮方法に於いて、前記タグ位置検出過程は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納過程で分離したタグ情報と共に格納することを特徴とするデータ圧縮方法。

【請求項23】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法に於いて、

前記タグ情報と符号データとを分離するタグ情報分離過程と、
前記タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、
前記符号データから文字列及びタグ符号を含む文字列ストリームを復元した後に、前記タグ符号を前記タグ情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備えたことを特徴とするデータ復元方法。

【請求項24】請求項23記載のデータ復元方法に於いて、前記文字列復元過程は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を生成する辞書生成過程と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書の参照で元の文字列を復元する文字列比較過程と、
前記文字列比較過程により復元したタグ符号を、前記タグ情報格納過程で分離したタグ情報に置き換える文字列置換過程と、を備えたことを特徴とするデータ復元方法。

【請求項25】請求項23記載のデータ復元方法に於いて、更に、前記タグ情報格納過程で格納されたタグ情報の圧縮データを復元するタグ情報復元過程を設けたことを特徴とするデータ復元方法。

【請求項26】請求項23記載のデータ復元方法に於いて、更に、
復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書の参照

で元のタグ文字列を復元するタグ文字列比較過程と、を備えたことを特徴とするデータ復元方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、タグを含む構造化文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置及び復元装置並びにその方法に関し、特に構造化文書の文字列ストリームからタグ情報を分離して符号化と復元を行うためのデータ圧縮装置及び復元装置並びにその方法に関する。

【0002】

【従来の技術】近年、文字コード、画像データ等の様々な種類のデータがコンピュータで扱われている。さらに、インターネット・イントラネットの普及に伴い、電子メールや電子化文書が増加している。このような大量のデータは、データ中の冗長な部分を省いて圧縮することにより、記憶容量を減らしたり、短時間で遠隔地に送ることを可能にしている。

【0003】本発明の分野は、文字コードの圧縮に限らず、様々なデータに適用できるが、以下では、情報理論で用いられる呼称を踏襲し、データの1ワード単位を文字と呼び、データが任意のワードつながったものを文字列と呼ぶことにする。

【0004】最近では、コンピュータ上で扱う文書の形式を統一する動きがある。その中で、文書の作成を効率良く行うため、タグを用いて文書内容を部分的に区別して、予め見出しや段落などの複数の文書部品を作成し、各々の文書部品間の関係を定めて文書を構造化して編集することが試みられている。

【0005】このような文書に構造化の概念を取り入れた構造化文書の例としては、国際規格のODA (ISO 8613: Open Document Architecture) や、SGML (ISO8879: Standard Markup Language) の規格による構造化文書がある。またこのような構造化文書を用いた文書処理方法は、例えば特開平5-135054号のものがある。

【0006】SGMLによる構造化文書は、従来のテキスト処理システムとの親和性が高く、米国を中心に普及し、実用化されてきている。SGMLによる構造化文書は、予め文書構造の雛型が与えられ、文書構造は雛型の範囲に制限される。

【0007】図25はSGMLの構造化文書であり、SGML宣言200、文書型定義 (DTD: Document Type Definition) 202、及び文書実現値204の3つの部分からなる。このうち文書の構造を定義する雛型が文書型定義202であり、図26のように、章、節、タイトルなどの文書構造を定義している。

【0008】SGMLの構造化文書では、文書構造を表現するために、文書テキスト内にタグと呼ばれる識別子を用いて、文書テキストを区分する。図27はSGML

の構造化文書の具体例であり、例えば文書のタイトルの場合、「<TITLE> 発明 (考案) 明細書</TITLE>」で表現される。即ち、開始タグである「<TITLE>」と終了タグである「</TITLE>」で囲まれた文字が要素であり、この場合はタイトル内容「発明 (考案) 明細書」を表わす。

【0009】現在、公的機関を中心にSGMLを採用する例が増えてきている。特に米国では、国防総省が文書をSGMLで記述して納入することを義務づけている。日本においても特許庁のCD-ROM公報として、この構造化文書を採用している。また、インターネットで使われているWWW (World Wide Web) の記述形式として普及しているHTML (Hyper Text Markup Language) は、SGMLの一形態である。

【0010】このようなSGML等の構造化文書を圧縮する方法として、本願出願人は、特開平9-261072号の方法を提案している。

【0011】この方法では、タグ情報を有する構造化文書の文書データが入力された場合、文書型定義DTDなどで定義されているタグ情報を検出する。タグ情報が検出された場合、タグ情報は何ら変換せずに、そのまま出力する。さらにタグ情報を検出したことにより、タグ情報以外の入力文字列を符号化するモードに移行する。

【0012】この符号化の基本アルゴリズムは図28のようになる。まずステップS1で入力された文字又は文字列が予め登録した辞書の文字又は文字列と同一か否かを検索して比較し、同一であればステップS2で入力データを辞書の登録番号で符号化し、ステップS3で符号を出力する。

【0013】ステップS1で同一の登録文字又は文字列が検索できなかった場合は、ステップS5で元の入力文字又は文字列をそのまま出力する。このような処理をステップS5で入力文字列がなくなるまで繰り返す。

【0014】図27のSGML文書ファイルについて図28の符号化を行うと、図29の圧縮データファイルが得られる。この圧縮データファイルは、1つのファイル中に圧縮されていないタグ情報の部分と圧縮されたテキスト文書の部分とが混在する形式となる。

【0015】

【発明が解決しようとする課題】文書テキストを圧縮する方法は、膨大なデータ量である文書テキストを実用に耐え得るデータ量に圧縮することができ、電子化された文書テキストを実現する上で非常に有用な技術である。

【0016】しかしながら、図29のような構造化文書の圧縮データファイルにあつては、ファイル中のタグ情報を検索する場合、タグ情報は圧縮されない部分として圧縮された文書データの中に混在しており、ファイル全体をメモリ上に展開して必要とするタグ情報を検索しなければならぬ。また圧縮部分となる本文中のキーワードを検索したい場合にも、同様にファイル全体をメモリ上に展開して処理する必要がある。

【0017】このため、構造化文書の圧縮データファイルから必要とする文書を検索あるいは入手するために、文書としては不要部分の読み込みが必要となり、データ伝送量が増加して読込みに時間がかかり、また大きなメモリ領域とディスク容量の確保が必要となる問題がある。

【0018】本発明の目的は、タグ情報を含む構造化文書の圧縮データにつき、文書を検索あるいは読込み時間の短縮とメモリやディスク容量の増加を最小限とするためのデータ圧縮装置及び復元装置並びにその方法を提供することにある。

【0019】

【課題を解決するための手段】図1は本発明の原理説明図である。

(圧縮) まず本発明は、図1(A)のように、タグを含む文書で構成された文字列ストリーム20から符号データを生成するデータ圧縮装置を対象とする。このデータ圧縮装置として本発明は、文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部10と、タグ情報分離部10でタグが分離された文字列ストリームの位置に識別のためにタグ符号24を配置するタグ符号置換部12と、タグ符号置換部12から出力されたタグ符号を含む文字列ストリーム22を符号化して符号ストリーム84を出力する文字列符号化部14とを設けたことを特徴とする。

【0020】このような本発明のデータ圧縮装置によれば、タグを含む構造化文書の文字列ストリームにつき、タグ情報と本文(文字列)とを分離し、少なくとも本文を符号化することで高い圧縮率を実現し、分離したタグ情報を検索することで、検索が高速化できる。

【0021】例えば、圧縮データファイルの中の本文から分離されたタグ情報を検索し、一致するタグ情報が検索できたら、復元した本文中のタグ符号を検索したタグ情報までの数だけ読み飛ばすことで、目標とする文書の先頭に容易に到達することができる。

【0022】タグ符号置換部12は、タグが分離された文字列ストリームの位置に、所定の固定符号をタグ符号として配置する。タグ符号として固定符号を使用することで、本文中のタグ位置の検索が簡単にできる。

【0023】またタグ符号置換部12は、タグが分離された文字列ストリームの位置に、タグ情報分離部10で分離されたタグの出現順序を示すタグ符号を配置する。このようにタグ符号に出現順序の情報を持たせることで、タグ情報に基づく本文検索の高速化と信頼性が高められる。

【0024】データ圧縮装置は、更に、タグ情報分離部10で分離されたタグ情報を格納するタグ情報格納部78と、文字列符号化部14で生成された符号データを格納する符号格納部80と、タグ情報格納部78に格納されたタグ情報と符号格納部80に格納された符号データ

を選択して出力する符号切替部82と設ける。このように分離したタグ情報と本文の符号データを個別に格納することで、圧縮データの検索や転送要求に対する管理を容易にする。

【0025】文字列符号化部14は、圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部18と、タグ符号置換部12からの文字列ストリーム22の中の部分文字列と辞書格納部18の登録文字列との比較により、登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部16とを備える。

【0026】この文字列符号化部14による符号化処理は、スペースで区切られない単語構造をもつ言語の文字コードで作成された文書データの圧縮に効果があり、スペースで区切られない単語構造をもつ言語としては、例えば日本語、中国語、ハングル語等がある。

【0027】日本語を例にとると、日本語の単語に関する(株)日本電子化辞書研究所(EDR)の研究がある(横井、木村、小泉、三吉、「表層レベルにおける電子化辞書の情報構造」、情報処理学会論文誌、Vol.37, No.3, p.333-344, 1996)。

【0028】この研究結果では、日本語を構成する形態素、即ち単語の品詞を集計している。単純に単語を品詞類に分けて登録すると136,486個となり、17ビット(最大262,143個)の符号で表わすことができる。

【0029】また新世代コンピュータ技術開発機構(ICOI)で作成した日本語単語辞書を構成する約13万語の単語ごとに構成する文字数を検出し、その分布を求めた結果、全登録単語の1/2以上の7万語が2文字で構成され、平均文字数は2.8文字(44.8ビット)であることも判明している。

【0030】図1(A)の辞書格納部18は、日本語の辞書として実用的な例えば約13万語の単語を、各単語に例えば17ビットの固定長の文字列符号を割り当てた辞書を作成して格納し、非圧縮データの部分文字列に一致する辞書の登録文字列を検索して17ビットの固定長符号を文字列符号として割り当てて出力することで、文書データの大小に関わらず、実質的にデータ量を半分以下に圧縮することができる。

【0031】本発明のデータ圧縮装置は、タグ情報分離部10で分離したタグ情報を圧縮するタグ情報圧縮部を設ける。タグ情報は、単独のタグと、タグと文字列の組合せを含むが、このタグ情報圧縮部は、タグと文字列を区別することなく一括してタグ情報を圧縮する。この圧縮は例えばLZ77、LZ78、算術符号化等のアルゴリズムを使う。

【0032】本発明のデータ圧縮装置は、タグ情報中の日本語等のスペースで区切られない言語の文字列を対象に、本文の文字列符号化部と同じ符号化を行ってタグ情

報を圧縮する。即ち、本発明のデータ圧縮装置は、圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、タグ情報分離部10で分離したタグ情報に含まれる文字列ストリームの部分文字列とタグ辞書格納部の登録文字との比較により、登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部とを備えたことを特徴とする。

【0033】このように分離したタグ情報についても圧縮することで、文字列符号化部14による本文の圧縮と併せて文書ファイル全体を高圧縮できる。

【0034】本発明のデータ圧縮装置に於いて、更に、文字列符号化部14で生成した符号データのタグ位置を検出するタグ位置検出部を設け、タグ情報格納部78にタグ情報分離部10で分離したタグ情報と共にタグ位置検出部で検出したタグ位置の指定情報を格納する。この場合、タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出してタグ情報格納部にタグ情報と共に格納する。

【0035】このように分離したタグ情報に、圧縮した本文の対応するタグ符号の位置を示す文書先頭又は特定タグからのデータ量（バイト数）が位置指定情報として格納されているため、タグ情報から必要とするタグを検索したい場合、直ちに本文に圧縮データ中の対応するタグ符号の位置が特定でき、必要とする本文のランダムアクセスが効率良くできる。

【0036】（復元）本発明は、タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置を対象とする。

【0037】このデータ復元装置として本発明は、図1(B)のように、符号ストリーム56からタグ情報と符号データとを分離するタグ情報分離部60と、タグ情報分離部60で分離したタグ情報を格納するタグ情報格納部62と、符号データから文字列及びタグ符号を復元した後に、タグ符号をタグ情報格納部62のタグ情報に置き換える文字列復元部64とを備えたことを特徴とする。

【0038】文字列復元部64は、図1(A)の文字列符号化部14の逆の操作を行うもので、復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部65と、符号ストリームから復元単位となる文字列の符号を分離して辞書格納部65の参照で元の文字列を復元する文字列比較部66と、文字列比較部66により復元したタグ符号をタグ情報格納部68のタグ情報に置き換える文字列置換部68とを備える。

【0039】本発明のデータ復元装置は、データ圧縮装置側でタグ情報をLZ77、LLZ78等で圧縮してい

る場合、タグ情報格納部62に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設ける。

【0040】また本発明のデータ復元装置は、データ圧縮装置側でタグ情報の文字列を符号化している場合、復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、タグ情報分離部60により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、タグ辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部とを備える。

【0041】本発明は、更に、タグ情報を含む構造化文書の圧縮方法及び復元方法を提供する。本発明によるタグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法は、文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を備える。

【0042】また本発明によるタグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法を提供する。この復元方法は、タグ情報と符号データとを分離するタグ情報分離過程と、タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、符号データから文字列及びタグ符号を復元した後に、タグ符号をタグ情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備える。データ圧縮方法及び復元方法の詳細は、装置の場合と同じになる。

【0043】

【発明の実施の形態】図2は本発明のデータ圧縮装置の第1実施形態のブロック図である。図2において、第1実施形態のデータ圧縮装置は、タグ情報分離部10、タグ符号置換部12及び文字列符号化部14で構成される。文字列符号化部14には文字列比較部16と辞書格納部18が設けられる。

【0044】タグ情報分離部10は、例えば図27に示したSGML日本語文書ファイルから読み出した文字列ストリーム20を入力し、入力した文字列ストリーム20に含まれるタグを識別し、識別したタグを分離してタグ情報ストリーム28として出力する。

【0045】タグ符号置換部12は、タグ情報分離部10でタグ情報が分離された文字列ストリームのタグ位置に予め定めたタグ符号を配置し、タグ符号配置済みの文字列ストリーム22を文字列符号化部14に供給する。文字列符号化部14はタグ符号置換部12により配置されたタグ符号を含む文字列ストリーム22を符号化し、

符号ストリーム26を出力する。

【0046】図3は、図2のタグ情報分離部10の詳細であり、タグ符号置換部12と共に示している。タグ情報分離部10は、タグ比較部30、タグ識別規則格納部32及び出力切替部34で構成される。タグ識別規則格納部32にはSGML文書における文書型定義DTDから得られたタグ情報の識別規則が格納されている。

【0047】タグ比較部30は文字列ストリーム20を入力し、タグ比較部30でタグ識別規則格納部32の識別規則と比較し、タグ情報識別により比較出力が得られると、出力切替部34を文字列ストリーム22の出力からタグ情報ストリーム28の出力に切り替え、識別したタグ情報をタグ情報ストリーム28として出力する。

【0048】同時にタグ符号置換部12にタグ情報識別に基づく比較出力を行い、タグ符号置換部12に予め設定されているタグ符号24を、出力切替部34から出力が断たれたタグ情報の位置に挿入配置する。タグ符号置換部12により文字列ストリーム22のタグ情報の位置に配置されるタグ情報24としては、例えば16進の固定符号「0x0000」を使用する。

【0049】図4は、SGML日本語文書ファイルから読み出した文字列ストリーム20の図2のデータ圧縮装置による圧縮処理の説明図である。

【0050】図2のタグ情報分離部10に対する文字列ストリーム20として入力されるSGML日本語文書ファイル35は、図3のタグ情報分離部10に設けているタグ比較部30でタグ識別規則格納部32に格納しているタグ識別規則と比較され、例えば先頭の「<TITLE>」の「<TITLE>」がタグ情報として識別され、このタグ情報はタグ情報ファイル36の先頭位置のように分離される。

【0051】またタグ情報の分離と並行して、SGML日本語文書ファイル35のタグ情報を分離した位置に、16進の固定符号「0x0000」を用いたタグ符号が挿入配置され、このタグ情報のタグ符号への置換によりタグ置換済み日本語文書ファイル38の文字列ストリームが生成される。

【0052】分離されたタグ情報ファイル36の内容となるタグ情報ストリームはそのまま出力される。またタグ置換済み日本語文書ファイル38の内容となる文字列ストリームは、文字列符号化部14により符号化されて符号ストリーム26として出力される。

【0053】図5は、図27のSGML日本語文書ファイルの文字列ストリーム20を図2のデータ圧縮装置に入力して、タグ符号置換部12によりタグ情報を固定タグ符号に置換して得たタグ置換済み日本語文書ファイル38である。このタグ置換済み日本語文書ファイル38では、図27のSGML日本語文書ファイルにおけるタグ情報がそれぞれ「(タグ符号)」に置き換えられている。

【0054】図6は図27に示すSGML日本語文書ファイルの文字列ストリームから分離したタグ情報のタグ情報ファイル36である。このタグ情報ファイル36には、入力した文字列ストリームに含まれているタグ情報が順番に分離されて格納されている。

【0055】図5のタグ置換済み日本語文書ファイル38の内容となるタグ置換済みの文字列ストリーム22は、図2の文字列符号化部14で符号化され、圧縮された符号ストリーム26として出力される。

【0056】図7は、タグ符号としてタグ情報の出現順序を示す順序タグ符号を使用した場合のタグ置換済み日本語文書ファイル38である。このタグ情報の出現頻度を表わす順序タグ符号としては、例えばタグの出現順序に応じて16進で「0x001, 0x002, 0x003, ...」等のように、一義に対応する順序タグ符号を使用すればよい。

【0057】この出現順序を示す順序タグ符号を使用した場合には、図7のように日本語文字列データの中に置換されたタグ符号自体が「(タグ符号1), (タグ符号2), (タグ符号3), ...」のように、文書先頭からの出現順序を表わしている。このため図6のように分離されたタグ情報の検索で図7の文書ファイル中の対応するタグ符号の位置を特定する際に、本文中の検索位置を簡単且つ確実に特定できる。

【0058】例えば図6で5行目のタグ情報「<SECTION>請求項目の範囲</SECTION>」の文書ファイル中の位置を知りたい場合には、このタグ識別情報は先頭から5番目に出現していることから、出現順序が5番目となる「(タグ符号5)」の位置を検索することで、簡単に特定できる。

【0059】図8は、図2のデータ圧縮装置による圧縮処理のフローチャートである。まずステップS1で、入力文書の文字列ストリーム20からタグ情報分離部10によってタグ情報を分離して出力する。続いてステップS2で、入力文書の文字列ストリーム20中のタグのあった位置にタグ符号置換部12によって識別のためのタグ符号を挿入する。

【0060】続いてステップS3で、タグ配置済みの文字列ストリーム中の文字列を文字列符号化部14に設けている文字列比較部16で辞書格納部18内の対応する登録番号を符号として割り当て、符号ストリーム26を出力する。このステップS1～S3の処理を、ステップS4で文字列ストリームの入力終了するまで繰り返す。

【0061】次に図2の文字列符号化部14に設けた文字列比較部16と、辞書格納部18によるタグ置換済み文字列ストリーム22の符号化処理を説明する。

【0062】図2の文字列符号化部14に設けた文字列比較部16は、辞書格納部15の参照により、単語を構成する文字列ごとに予め定めた所定の文字列符号を割り

当てる符号化を行う。

【0063】まず文字列比較部16で圧縮対象とする文書データとして、例えば日本語文書データを例にとると、日本語文書データの場合、1文字は2バイトのワードデータで構成されており、文書中の単語はスペースで区切られない構造を持っている。また日本語文書データは、1回の圧縮に使用する文書単位に入力しており、これはキロバイトオーダからメガバイトオーダの適宜のサイズの文書が入力される。

【0064】文字列比較部16は日本語文書データの文字列を先頭から順番に入力し、辞書格納部15に予め登録されている単語単位の登録文字列と一致するかどうかを検出する。文字列比較部16で入力文字列に一致する登録文字列が検出されると、辞書格納部15の一致検出された登録文字列に対応して予め登録されている文字列符号を読み出して割り当て、この文字列符号を文字列置換部18に出力する。

【0065】ここで日本語文書データの文字列を単語単位に文字列符号に変換するための辞書格納部15を説明する。

【0066】図9は、(株)日本電子化辞書研究所(EDR)が研究成果として発表した、日本語を構成する形態素の品詞に関する集計結果である。この集計結果を見ると、単語数に対応する形態素数は136、486個であり、この単語の数を2進数で表現すると、最大表現数が262、143個となる17ビットの符号で表すことができる。

【0067】これに対し、新世代コンピュータ技術開発機構(ICOT)で作成した約13万語の単語を有する日本語辞書から単語を構成する文字数を検出して分布を求めた結果、全登録単語の1/2以上の7万語が2文字で構成されており、平均文字数は2.8文字となっている。この平均文字数2.8文字をビット数で表すと、 $2.8 \text{ 文字} \times 2 \text{ バイト} = 5.6 \text{ バイト} \times 8 \text{ ビット} = 44.8 \text{ ビット}$ となる。

【0068】そこで本発明にあっては、図9の136、486個の単語を表現する17ビットの文字列符号を予め割り当て、入力した日本語データの文字列を単語単位に17ビットの文字列符号に変換する符号化を行うことで、実質的にデータ量を半分に以下に圧縮することができる。

【0069】図10は、図2の辞書格納部15の辞書構造の実施形態である。図2の辞書格納部15に格納された辞書は、先頭文字格納部40と従属文字列格納部42

$$K = (N \cdot X - A1) / M$$

但し、X：従属文字列格納部22の位置アドレス

N：一致検出された従属文字列の番号(1, 2, 3, …, N)

A1：従属文字列格納部の開始アドレス

の2階層構造を備える。先頭文字格納部40は、日本語文字「あ、い、う、え、お…」の文字コードをインデックスとしており、日本語の文字コードは2バイトデータであることから、文字コード44としては、16進数で「0x0000」から「0xFFFF」の131,072種類の格納位置が割り当てられる。

【0070】この文字コード44は、図2の文字列比較部16で読み込んだ先頭文字を使用して、対応する文字コードの位置にアクセスする。文字コード24に続いては先頭アドレス46が格納される。先頭アドレス46は、例えば文字コード44の先頭文字「あ」を例にとると、先頭文字「あ」に続く従属文字列を格納した従属文字列格納部42の先頭アドレス「A1」を指定している。続いて従属文字列の個数48が設けられる。例えば先頭文字「あ」にあつては、従属文字列個数48としてN1=4個が格納されている。

【0071】従属文字列格納部42は、先頭文字格納部40の先頭文字の文字コード44に対応して格納された先頭アドレス46で先頭位置が指定され、この先頭位置から従属文字列格納部48で指定された個数の格納位置に従属文字列が格納されている。例えば先頭文字「あ」に対応した先頭アドレス46のアドレスA1から従属文字列個数48のN1=4個となる4つの格納位置が、対象とする従属文字列格納領域として指定される。

【0072】この従属文字列格納部42は、先頭から従属文字列の長さ50、従属文字列52、及び17ビット表現される文字列コード(文字列符号)54が格納されている。例えば先頭アドレスA1には、長さL1で従属文字列「い」と、その文字列コードが格納されている。次の格納位置には長さL2の従属文字列「う」がその文字列コードと共に格納されている。

【0073】3番目の領域には長さL3の従属文字列「お」が文字列コードと共に格納されている。4番目の格納領域には長さL4で従属文字列が存在しないことを示す符号「NULL」が格納され、存在しないことを示す文字列コードが格納されている。即ち、この4番目の格納領域は先頭1文字だけの文字列コードの登録を表している。

【0074】ここで図10の従属文字列格納部42に文字列コード34は、単語個数に基づき1番から136、486番まで予め17ビットの文字列コードが割り当てられており、図10のように格納した場合の文字列コード(文字列符号)Kと位置アドレスXとの関係は、次式で表すことができる。

$$(1)$$

M：従属文字列格納部の格納バイト長

ここで、従属文字列格納部42の格納バイト長Mは、従属文字列の長さ50、従属文字列52、及び文字列コード54の合計長であることから、例えば次式で表すこと

ができる。

$$\begin{aligned}\text{格納バイト長}M &= \text{長さ} + \text{文字コード列} + \text{文字列コード} \quad (2) \\ &= 3\text{bit} + 96\text{bit} + 17\text{bit} \\ &= 116\text{bit} \\ &= 15\text{byte}\end{aligned}$$

ここでは格納可能な従属文字列を最大6文字とすることで、従属文字列52に96ビットを割り当てた場合を例にとっている。もちろん、従属文字列の文字数は平均で2.8文字であることから、3文字(48ビット)以上とすれば十分な圧縮効果が得られる。この場合には、従属文字列格納部の1つの格納領域の格納バイト長MはM

$$X = M \cdot K + A1$$

但し、K : 文字列コード

A1 : 従属文字列格納部の開始アドレス

M : 復元側の従属文字列格納部の格納バイト長

この(3)式にあっては、復元側で使用する辞書における従属文字列格納部42の開始アドレスA1、即ちオフセットと、従属文字列格納部42の格納バイト長Mが定数として定まっていることから、復元しようとする文字列コードKを(3)式に代入することで、一義的に復元しようとする文字列を格納した辞書位置(位置アドレス)Xを算出することができる。

【0076】図11は図10の辞書構造を持った辞書格納部15による図2の文字列比較部16による符号化処理のフローチャートである。

【0077】まずステップS1で、文字列比較部16に読み込んだ文字列の先頭文字の位置Pにポインタを移動し、ステップS2で先頭文字位置Pの文字コードが示す図10の文字コード44に対応した先頭文字格納部40のテーブルを参照する。この先頭文字格納部40のテーブル参照により、ステップS3で従属文字列格納部42の先頭アドレス46と従属文字列の個数48を取得する。

【0078】続いてステップS4で、従属文字列格納部42の先頭アドレスの先頭データから従属文字列の長さ50の長さデータLを取得する。次にステップS5で、先頭文字位置Pから従属文字列の長さデータLに基づくL文字を抽出し、抽出したL文字を従属文字列格納部42の従属文字列52の登録文字列と比較して、一致するか否かを判断する。

【0079】登録した従属文字列と一致すれば、ステップS3に進み、次の文字列コード54を読み出し、一致検出した文字列に文字列比較部16で割り当てて出力し、次のステップS9で先頭文字位置Pを従属文字列の文字数Lだけ移動した位置Pにポインタを更新する。そしてステップS12で非圧縮データの処理が済んでいなければ、再びステップS2に戻り、更新した先頭文字位置Pについて同様な処理を繰り返す。

【0080】一方、ステップS5で従属文字列格納部42の登録従属文字列と一致しなかった場合には、従属文

＝12バイトになる。

【0075】前記(1)式から算出される17ビットの文字列コードKを使用した場合、復元の際には文字列コードKの値から格納位置(アドレス)Xを次式で算出すればよい。

$$(3)$$

文字数Nが未了か否かをチェックし、未了であればステップS7に戻り、従属文字列格納部42の先頭アドレスの次の格納領域から従属文字列の長さデータLを取得し、ステップS5で再度、先頭文字位置PからL文字の従属文字列を抽出して従属文字列格納部42の登録従属文字列と一致するか否かを比較する。

【0081】ステップS5～S7の繰返しによる登録個数Nの従属文字列の全てについて比較処理を行っても一致しなかった場合には、ステップS6で従属文字列の個数Nの終了を判別して、ステップS10に進み、先頭文字1文字を表す未登録コードを送出する。そしてステップS11で先頭文字位置Pを文字数L＝1文字だけ移動した次の位置にポインタを更新し、ステップS12からステップS2に戻って、次の先頭文字位置Pからの処理を繰り返す。

【0082】図12は、図2のデータ圧縮装置から出力された符号処理部26及びタグ情報ストリーム28で構成される符号ストリームから文字列ストリームを復元するためのデータ復元装置の第1実施形態のブロック図である。

【0083】このデータ復元装置は、タグ情報分離部60、タグ情報格納部62、文字列復元部64で構成される。文字列復元部64は、符号列比較部66、辞書格納部70及び文字列置換部68を備える。

【0084】タグ情報分離部60は図2のデータ圧縮装置側から送られてきた符号ストリーム58を入力し、タグ情報と符号データとに分離し、タグ情報はタグ情報格納部62に格納し、符号データは符号ストリーム58として文字列復元部64に出力する。

【0085】文字列復元部64は符号列比較部66で辞書格納部70を用いて符号データから文字列及びタグ符号を復元した後に、文字列置換部68においてタグ符号をタグ情報格納部62に格納しているタグ情報に置き換えて、復元した文字列ストリーム70を出力する。

【0086】図13は、図12のデータ復元装置の復元処理のフローチャートである。まずステップS1で、タグ情報分離部60が入力文書に対応した符号ストリーム56からタグ情報を分離してタグ情報格納部62に格納

する。次にステップS2で、タグ情報が分離された符号ストリーム58の中の符号列を辞書格納部65内の登録番号と比較照合し、一致する登録番号で格納している文字または文字列に変換する。

【0087】続いてステップS3で、復元された文字列に含まれているタグ符号をタグ情報格納部62に格納しているタグ情報の格納順に順次置換し、復元した文字列ストリーム70として出力する。これらのステップS1～S3の処理を、ステップS4で符号ストリーム56の入力が終了するまで繰り返す。

【0088】図12の文字列復元部64に設けた符号列

$$\begin{aligned} \text{格納バイト長}M &= \text{先頭文字} + \text{長さ} + \text{文字コード列} & (6) \\ &= 16\text{bit} + 3\text{bit} + 96\text{bit} \\ &= 115\text{ビット} \\ &= 15\text{byte} \end{aligned}$$

から判明しているため、次式から文字列コードKに対応

$$X = M \cdot K + A1$$

但し、K : 文字列コード

A1 : 文字列格納位置の開始アドレス

M : 格納バイト長

このようにして分離した文字列コードKから辞書格納位置を示す位置アドレスXを求めて参照することで、対応する先頭文字及び従属文字列を組み合わせた文字列を復元することができる。

【0090】このような図2のデータ圧縮装置及び図12のデータ復元装置により、図27に示したSGML日本語文書ファイルの文字列ストリームは、図6のようなタグ情報と図5のようなタグ情報をタグ符号に置き換えた文字列ストリームに分離され、この実施形態にあっては、タグ符号に置換済みの文字列ストリームを符号化することで文書ファイルの本文に相当する部分を高い圧縮率の圧縮ファイルに変換できる。

【0091】また図6のように分離されたタグ情報について、キーワードを使用して検索し、キーワードに一致するタグ情報が得られたならばタグ情報の出現位置が何番目かを検出し、これによって図5のタグ符号置換済みの本文の文書ファイルに含まれているタグ符号の出現位置を検索することで、タグ情報の検索結果に対応した文書位置の特定による読出し等が容易にできる。

【0092】図15は、本発明のデータ圧縮装置の第2実施形態であり、この実施形態にあっては、図2の第1実施形態に加えてタグ情報格納部78と符号格納部80を設けたことを特徴とする。

【0093】タグ情報格納部78にはタグ情報分離部10により文字列20から分離されたタグ情報が格納される。これによってタグ情報格納部78には、例えば図6のようなタグ情報ファイル36が格納される。また符号格納部80は文字列符号化部14に設けられており、タグ符号置換部12により分離したタグ情報にタグ情報を挿入したタグ置換済み文字列ストリーム22につき、図

比較部66は、辞書格納部65の参照により、図3のデータ圧縮装置で符号化された符号化ストリームから元の文字列を復元する。

【0089】図14は、図12の文字列辞書格納部70の辞書構造である。この文字列辞書格納部70にあっては、先頭文字72、従属文字列長さ74及び従属文字列76を、図10の辞書構造に示した従属文字列格納部42の17ビットの文字列コード54の順番に格納している。このため符号化比較部40にあっては、復元に使用する文字列辞書格納部42の格納バイト長Mが

した位置アドレスXを算出することができる。

$$(7)$$

11の符号化処理により生成された符号データが格納される。

【0094】このようなタグ情報格納部78及び符号格納部80に加え、出力段に符号切替部82が設けられる。符号切替部82は、タグ情報格納部78に格納されたタグ情報と符号格納部80に格納された符号データを、例えば順番に選択して符号列ストリーム84として出力する。

【0095】図16は、図15のデータ圧縮装置の圧縮処理のフローチャートである。この圧縮処理は、ステップS1で、入力文書の文字列ストリーム20からタグ情報分離部10でタグ情報を分離し、タグ情報格納部78に格納する。次にステップS2で、文字列ストリーム20の中のタグのあった位置にタグ符号置換部12によって識別用のタグ符号を挿入する。

【0096】次にステップS3で、タグ符号の置換が済んだ文字列ストリーム22の文字列を文字列符号化部14の文字列比較部16に入力し、辞書格納部18内の辞書構造の対応する登録番号に変換する。このようなステップS1～S3の処理を、ステップS4で文字列ストリームの入力が終了するまで繰り返す。

【0097】文字列ストリームの入力が終了するとステップS5に進み、分離したタグ情報とタグ符号に変換して符号化した符号ストリームを、例えばタグ情報格納部78と符号格納部80から順番に読み出して符号ストリーム84として出力する。図15のデータ圧縮装置から出力された符号ストリーム84は、図12に示したデータ復元装置に入力することで文字列ストリームを復元することができる。

【0098】図17は、本発明のデータ圧縮装置の第3実施形態であり、この実施形態にあっては文字列ストリームから分離したタグ情報を圧縮するようにしたことを特徴とする。

【0099】図17において、このデータ圧縮装置は、図15の第2実施形態におけるタグ情報分離部10とタグ情報格納部78の間に、新たにタグ情報圧縮部86を設けている。タグ情報圧縮部86は、タグ情報分離部10において入力した文字列ストリーム20から分離したタグ情報を圧縮対象の文字列ストリームとして圧縮してタグ情報格納部78に格納する。

【0100】タグ情報圧縮部86による圧縮処理は、タグ情報にはタグと日本語文字列が含まれ、これらを一括して圧縮することから、LZ77、LZ78、算術符号化等の圧縮アルゴリズムを使用する。

【0101】タグ情報分離部10、タグ符号置換部12、文字列符号化部14は、図15の第2実施形態と同じである。

【0102】図18は、図17のデータ圧縮装置による圧縮処理の説明図である。SGML日本語文書ファイル35の内容となる文字列ストリーム20は、タグ情報分離部10によってタグ情報ファイル36の内容となるタグ情報に分離される。このタグ情報はタグ情報圧縮部86により圧縮した後、タグ情報格納部78の格納を介して出力する。

【0103】またSGML日本語文書ファイル35の内容となる文字列ストリーム20から分離したタグ情報の位置には、タグ符号置換部12によって固定タグ符号または出現順序を示す順序タグ符号が挿入配置され、タグ置換済み日本語文書ファイル38の内容となる文字列ストリーム22が文字列符号化部14に出力され、文字列符号化により圧縮された符号データが符号格納部80による格納を介して出力される。

【0104】図19は、図17のデータ圧縮装置から出力された符号ストリーム90から文字列ストリームを復元する本発明のデータ復元装置の第2実施形態である。このデータ復元装置は、図12の第1実施形態に更に圧縮タグ格納部62とタグ情報復元部92を設けている。

【0105】タグ情報分離部60は、入力する符号ストリーム90に含まれる圧縮タグ情報を分離して圧縮タグ格納部62に格納する。圧縮タグ格納部62に格納された圧縮タグ情報はタグ情報復元部92により復元され、タグ情報格納部62に格納される。タグ情報復元部92はデータ圧縮側のLZ77、LZ78、算術復号化に対応した復元アルゴリズムを実行する。それ以外の構成は図15と同じになる。

【0106】図20は、本発明のデータ圧縮装置の第4実施形態であり、分離したタグ情報の中の日本語文字列を符号化により圧縮し、更に、分離したタグ情報に本文中の置き換えを行ったタグ符号の位置を示す位置指定情報を付加するようにしたことを特徴とする。

【0107】図20において、タグ情報分離部10、タグ符号置換部12、文字列比較部16、辞書格納部18を備えた文字列符号化部14、タグ情報格納部78及び

符号出力部82は、図15の第2実施形態と同じである。これに加えて図20の第4実施形態にあつては、新たにタグ文字列比較部94、タグ辞書格納部96及び符号量計測部98を設けている。

【0108】タグ文字列比較部94とタグ辞書格納部96は、タグ情報分離部10で分離したタグ情報に含まれる日本語文字列ストリームを文字列符号化部14と同様な符号化アルゴリズムで符号化してタグ情報を圧縮する。このため、タグ情報格納部96の辞書構成は図10と同じであり、先頭文字及び従属文字としてタグ情報に使用する日本語文字列が使用されている。またタグ文字列の符号化処理は図11のフローチャートに従って行う。

【0109】一方、図20のデータ圧縮装置に設けた符号量計測部98は、文字列符号化部14による本文の文字列ストリーム22、即ちタグ符号の置換が済んだ文字列ストリーム22を対象とした符号化による符号データについて、文字列ストリームの先頭から置換済みの各タグ符号までの符号量を計測し、この各タグ符号までの符号量の計測結果を、タグ情報格納部78に格納する文字列ストリームから分離した各タグ情報のそれぞれに符号位置情報として付加して格納する。

【0110】符号量計測部98によるタグ符号で置換されたタグ情報の位置を示す位置指定情報としては、文字列ストリームの先頭からの符号量以外に、文字列ストリームの中に特定のタグ情報からの後続する各タグ情報までの符号データの符号量としてもよい。

【0111】図21は、図20の第4実施形態における圧縮処理の説明図である。SGML日本語文書ファイル35の内容となる文字列ストリームを入力して、タグ情報の分離によるタグ情報ファイル36の生成及びタグ情報をタグ符号に置換したタグ置換済み日本語文書ファイル37の生成は、図15の第2実施形態と同じである。

【0112】これに加えて、分離されたタグ情報ファイル36のタグ情報に含まれている日本語文字列であるタグ文字列を、タグ辞書格納部96を用いて符号化して圧縮することで出力している。

【0113】図22は、タグ情報格納部78に格納されたタグ情報ファイルの具体例であり、図27に示したSGML日本語文書ファイルから分離したタグ情報を例にとっている。このタグ情報ファイル36には、左側のインデックス01～13に対応した各タグに対応して、右側に図21のタグ置換済み日本語ファイル37の文字列データの符号データの先頭からの符号量（バイト量）DL1～DL13が位置指定情報106としてそれぞれ格納されている。

【0114】図23は、図20の第4実施形態による圧縮処理のフローチャートである。まずステップS1～S4は図8と同じであり、タグ情報分離部10で文字列ストリーム20から分離したタグ情報をタグ情報格納部7

8に格納し、またタグ符号置換部12により分離したタグ情報の位置にタグ符号24を挿入配置した文字列ストリーム22を文字列符号化部14で符号化して符号データを符号格納部80に格納する。

【0115】次のステップS4にあっては、符号量計測部98が文字列符号化部14で置換済みのタグ符号を符号化する際に、例えば文字列ストリームの先頭からの符号量DLを計測し、既にタグ情報格納部78に格納されているタグ情報に図22の位置指定情報106として計測した符号量DLを格納する。

【0116】このようなステップS1～S4の処理を、ステップS5で文字列ストリームの入力終了するまで繰り返す。文字列ストリーム20の入力が終了すると、ステップS6で、タグ情報格納部78に分離したタグ情報中の文字列をタグ辞書格納部96内の辞書の対応するブロック番号に変換して符号データとする符号化処理をタグ文字列比較部94で行い、タグ情報格納部78に格納する。その結果、タグ情報格納部78の格納内容は図22の圧縮タグ情報ファイル36のようになる。

【0117】最後にステップS7で、タグ情報格納部78に分離して符号化した符号量付きのタグ情報と符号化部80に格納した符号データを符号切替部82より例えば順番に選択出力し、符号ストリーム100として外部に供給する。

【0118】ここで、図23の圧縮処理にあっては、ステップS1～S4のタグ情報の分離と置換、更には圧縮した符号量の計測処理とその後の分離したタグ情報の符号化処理を、時間的に分けて処理しているが、両者を並行して処理するようにしてもよいことはもちろんである。

【0119】図24は、図20のデータ圧縮装置から出力された符号ストリーム100から文字列ストリームを復元する本発明のデータ復元装置の第3実施形態である。

【0120】図24において、タグ情報分離部60、圧縮タグ格納部92、タグ情報格納部62、文字列復元部64は図19の第2実施形態と同じであり、これに加えて図24の第3実施形態にあっては、タグ文字列復元部102とタグ復元辞書格納部104を新たに設けている。

【0121】タグ復元辞書格納部104は、図13の辞書構造と同じものが使用され、格納している文字がタグに使用している日本語文字列となっている。タグ情報分離部60は、図20のデータ圧縮装置側から供給される符号ストリーム100から、図22の圧縮タグ情報ファイル36の内容に示すようなタグ情報ストリームを分離し、圧縮タグ格納部92に格納する。

【0122】圧縮タグ格納部92に格納された圧縮タグ情報は、タグ文字列復元部102によるタグ復元辞書格納部104のタグ文字列の符号による辞書番号の参照で

対応する日本語文字列に復元され、復元した日本語文字列を含むタグ情報をタグ情報格納部62に格納する。

【0123】一方、タグ情報分離部60は、圧縮タグ情報ストリームに続いて送られてくる文書本文の符号ストリームを文字列復元部64に供給し、符号列比較部66で取り出した符号による辞書格納部65の辞書番号の参照で対応する文字または文字列を復元し、符号置換部68に出力する。

【0124】符号置換部68は、復元した文字列の中のタグ符号を認識し、その出現順に従ってタグ情報格納部62に格納している復元済みのタグ情報を格納順に取り出し、タグ符号と置換し、復元した文字列ストリームを出力する。

【0125】ここで、圧縮タグ格納部92には、符号ストリーム100から分離した圧縮タグ情報ストリームの入力が終了した時点で、図22のように圧縮タグ情報ファイル36が格納されている。そこで圧縮タグ情報ファイル36について、特定のタグをキーワードとして検索を行い、一致するタグが得られたならば、これに対応する位置指定情報としての符号量DLを読み出し、図20のデータ圧縮装置に対し検索した符号量DLの位置からの符号データの転送要求ができる。これによって、データ復元側からデータ圧縮側に必要とするSGML日本語文書の部分的な圧縮本文データの転送による読込みが簡単にできる。

【0126】尚、本発明におけるデータ圧縮装置からデータ復元装置への伝送形態としては、インターネット等の通信回線でもよいし、光ディスクカートリッジや磁気ディスクカートリッジ等の書替可能な可搬媒体等の適宜の形態でよい。

【0127】また上記の実施形態は、タグ情報を分離し、分離したタグ情報の位置にタグ符号を置換した文字列ストリームの圧縮として、日本語固有の単語数に対応した固定長の文字列符号を割り当てた符号化を例にとっているが、これ以外のLZ77、LZ78、算術符号化等の圧縮を行うようにしてもよいことはもちろんである。

【0128】更に、本発明は、上記の実施形態の数値による限定は受けない。更に本発明は、その目的と利点を損なわない範囲の適宜の変形を含む。

【0129】

【発明の効果】以上説明してきたように本発明によれば、タグを含むSGML等の構造化文書の文字列ストリームにつき、タグ情報と本文（文字列）とを分離し、少なくとも本文を符号化することで高い圧縮率を実現し、また分離したタグ情報を検索することで圧縮された符号データ中の特定のタグ位置の読出しや検索を高速で処理することができる。

【0130】即ち、分離したタグ情報の順番と符号データ中に置換したタグ符号の順番は1対1に対応してお

り、タグ情報について特定のタグ情報を検索すること
で、その順番から符号データ中のタグ符号の位置が特定
でき、目標とする文書符号データの先頭位置に容易に到
達することができる。

【0131】この結果、タグを含むSGML等の構造化
文書について、高い圧縮率を保ちながら高速に圧縮及び
復元を行うことができる。

【図面の簡単な説明】

【図1】本発明の原理説明図

【図2】本発明のデータ圧縮装置の第1実施形態のブ
ロック図

【図3】図2のタグ情報分離部のブロック図

【図4】図2のデータ圧縮装置の処理手順の説明図

【図5】図4のタグをタグ符号に置換した本文ファイル
の説明図

【図6】図4の文字列ストリームから分離したタグ情報
ファイルの説明図

【図7】図4のタグを出現順序付きのタグ符号に置換し
た本文ファイルの説明図

【図8】図2のデータ圧縮装置の圧縮処理のフローチャ
ート

【図9】日本語文書に対する研究結果の説明図

【図10】図2の辞書格納部の辞書構造の説明図

【図11】図10の辞書構造を用いた図2の符号化処理
のフローチャート

【図12】図2のデータ圧縮装置からの符号ストリーム
を復元する本発明のデータ復元装置の第1実施形態のブ
ロック図

【図13】図12のデータ復元装置の復元処理のフロー
チャート

【図14】図12の辞書格納部の辞書構造の説明図

【図15】本発明のデータ圧縮装置の第2実施形態のブ
ロック図

【図16】図15のデータ圧縮装置の圧縮処理のフロー
チャート

【図17】本発明のデータ圧縮装置の第3実施形態のブ
ロック図

【図18】図17のデータ圧縮装置の処理手順の説明図

【図19】図17のデータ圧縮装置からの符号ストリー
ムを復元する本発明のデータ復元装置の第2実施形態の
ブロック図

【図20】本発明のデータ圧縮装置の第4実施形態のブ
ロック図

【図21】図20のデータ圧縮装置の処理手順の説明図

【図22】図21の符号量をタグに付加した図20のデ
ータ圧縮装置に格納されるタグ情報ファイルとタグ情報
ストリームの説明図

【図23】図20のデータ圧縮処理のフローチャート

【図24】図20のデータ圧縮装置からの符号ストリー

ムを復元する本発明のデータ復元装置の第3実施形態の
ブロック図

【図25】SGML文書の構造説明図

【図26】SGML文書の文書型定義DTDの具体例の
説明図

【図27】日本語文書を例にとったSGML文書ファイ
ルの説明図

【図28】SGML文書ファイルを圧縮する基本的な符
号化アルゴリズムのフローチャート

【図29】圧縮していないタグ情報の部分と圧縮した本
文部分が混在したSGML文書圧縮データファイルの説
明図

【符号の説明】

10：タグ情報分離部

12：タグ符号置換部

14：文字列符号化部

16：文字列比較部

18：辞書格納部

20：文字列ストリーム

22：タグ置換済み文字列ストリーム

24：タグ符号

26, 84, 90, 102：符号ストリーム

28, 100：タグ情報ストリーム

30：タグ比較部

32：タグ識別規則格納部

34：出力切替部

35：SGML日本語文書ファイル

36：タグ情報ファイル

38：タグ置換済み日本語文書ファイル

40：先頭文字格納部

42：従属文字列格納部

60：タグ符号分離部

62：タグ情報格納部

64：文字列復元部

66：符号列比較部

68：文字列置換部

70：辞書格納部

78：タグ情報格納部

86：タグ情報圧縮部

88：圧縮タグ情報ストリーム

92：圧縮タグ格納部

94：タグ情報復元部

95：符号量計測部

96：タグ文字列比較部

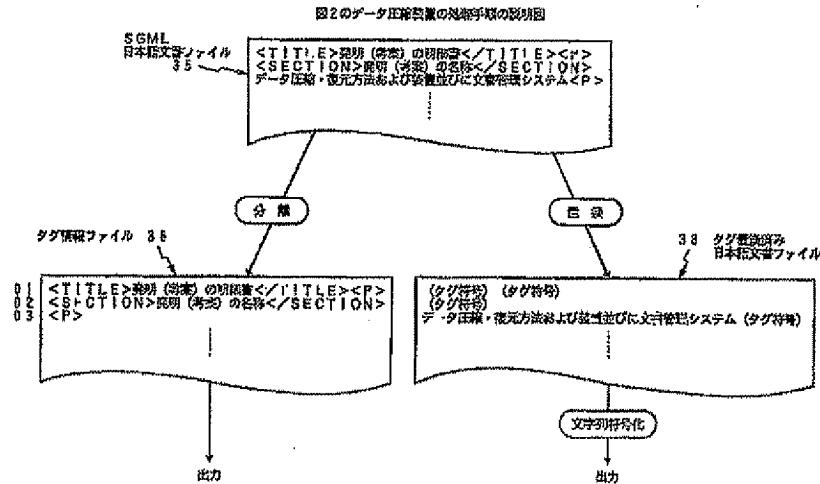
98：タグ辞書格納部

98：符号量計測部

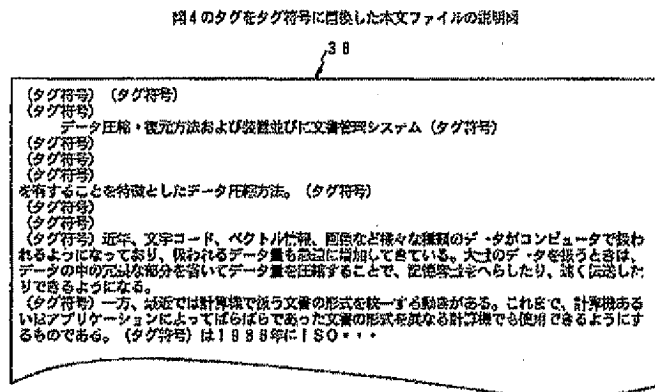
104：タグ文字列復元部

106：タグ復元辞書格納部

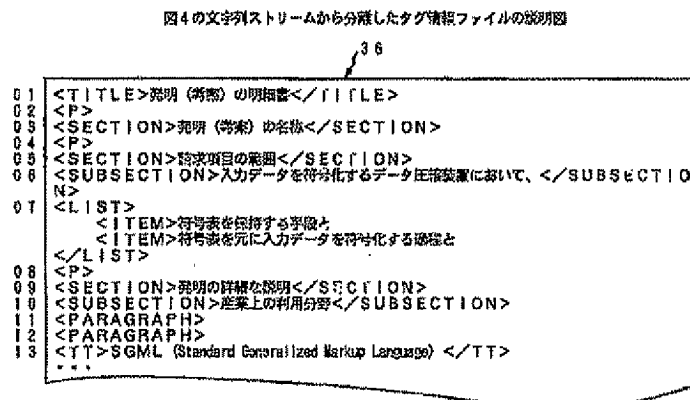
【图4】



【图5】



【图6】

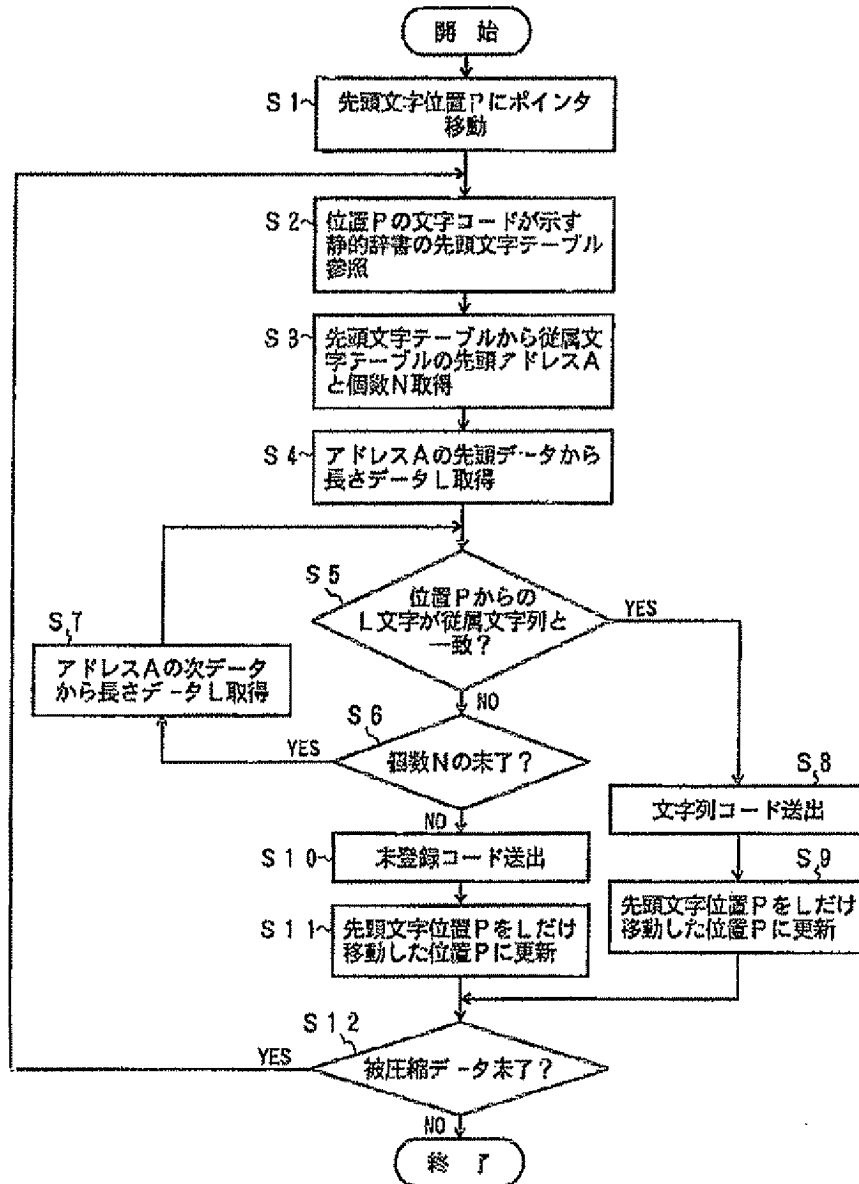


2.0.0	2.0.2	2.0.4
SGML 语言	DTD	文書實現值

SGML 文書

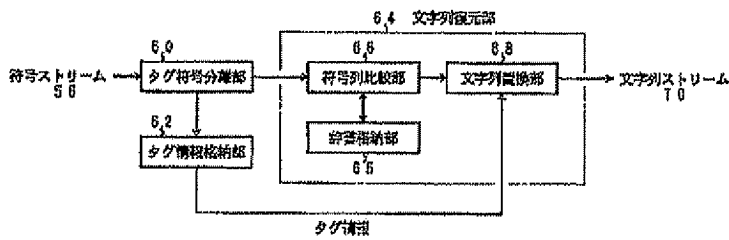
【 図 11 】

図 10 の辞書構造を用いた図 2 の符号化処理のフローチャート



【図12】

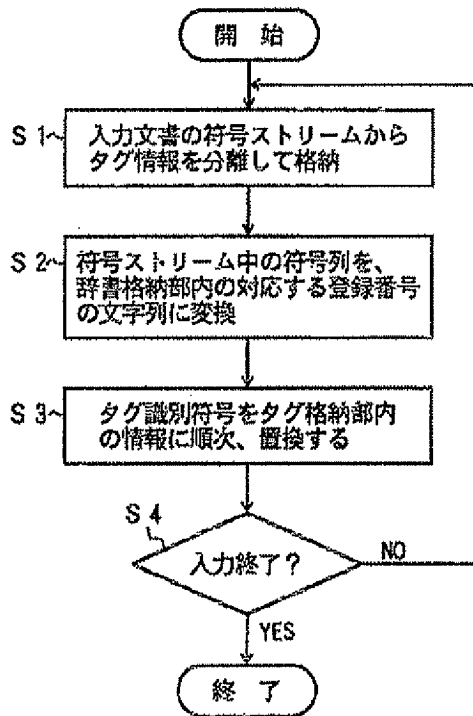
図2のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第1実施形態のブロック図



【図13】

【図26】

図12のデータ復元装置の復元処理のフローチャート



【図15】

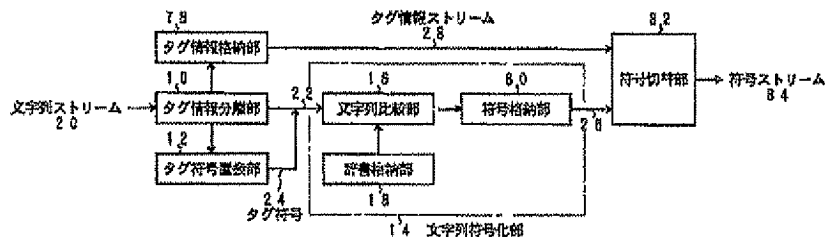
SGML文書の文書型定義DTDの具体例の説明図

```

*****
<!--=====Link Markup=====-->
<ENTITY % linkType "NAME">
<ENTITY % linkExtraAttributes
  "REL %linkType#IMPLIED
  REV %linkType#IMPLIED
  URN COAT#IMPLIED
  TITLE COAT#IMPLIED
  METHODS NAMES#IMPLIED"
>
<![% HTML Recommended [
  <ENTITY % A.content "(%text)*"
  --<H1> <a name="xxx">Heading</a></H1>
  is preferred to
  <a name="xxx"></H1>Heading</H1></a>
  ]]>
<ENTITY % A.content "(%heading | %text)*"
<ELEMENT A
  --%A.content-(A)>
  <ATTLIST A
    REF COAT#IMPLIED
    NAME COAT#IMPLIED
    %linkExtraAttributes;
    %SDAPDF;"<Anchor:#AttList>"
  <!--<A> Anchor:source/destination of link -->
  <!--<A NAME="..."> Name of this anchor -->
  <!--<A REF="..."> Address of link destination -->
  <!--<A URN="..."> Permanent address of destination -->
  *****

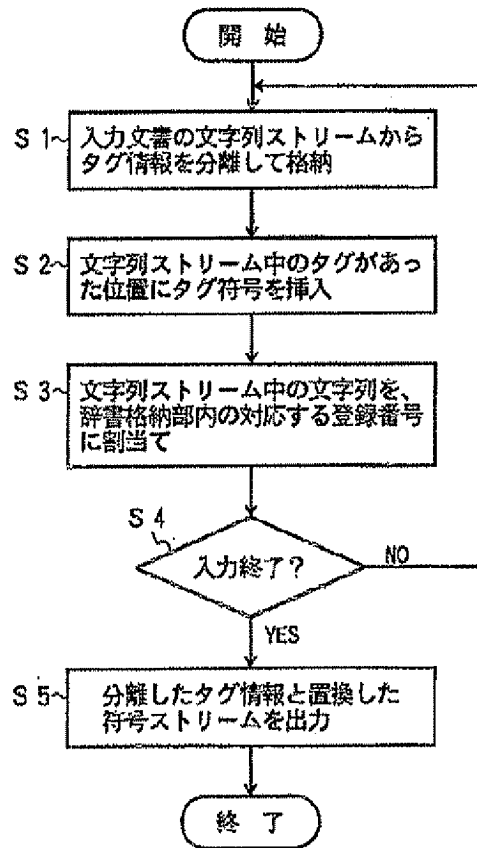
```

本発明のデータ圧縮装置の第2実施形態のブロック図



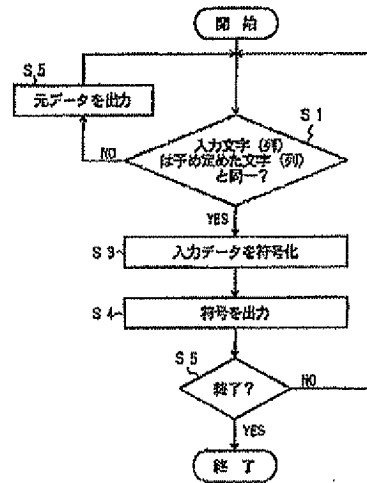
【図16】

図15のデータ圧縮装置の圧縮処理のフローチャート



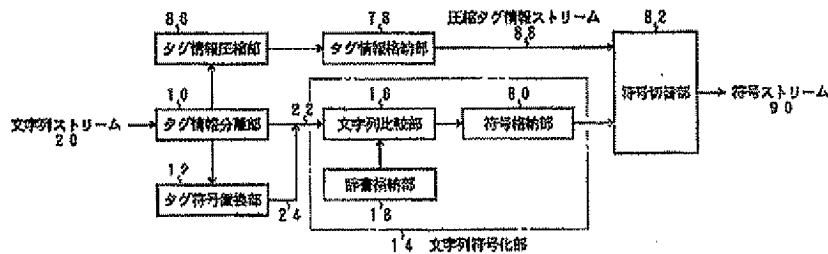
【図28】

SGML文書ファイルを圧縮する基本的な符号化アルゴリズムのフローチャート

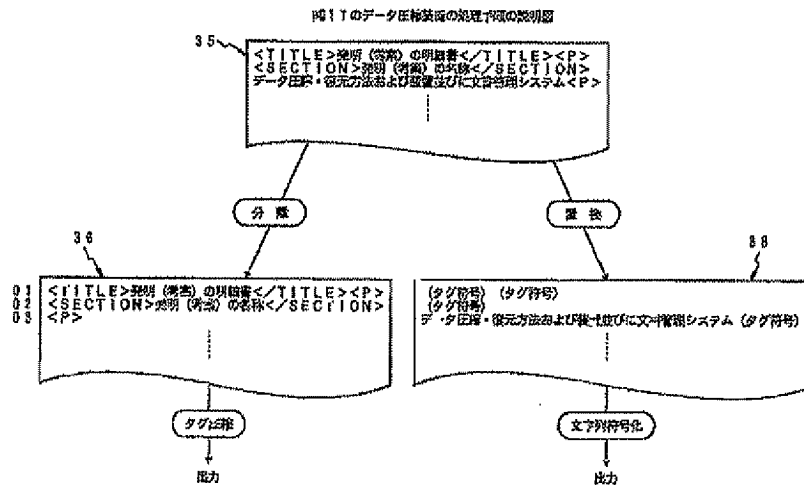


【図17】

本発明のデータ圧縮装置の第3実施形態のブロック図

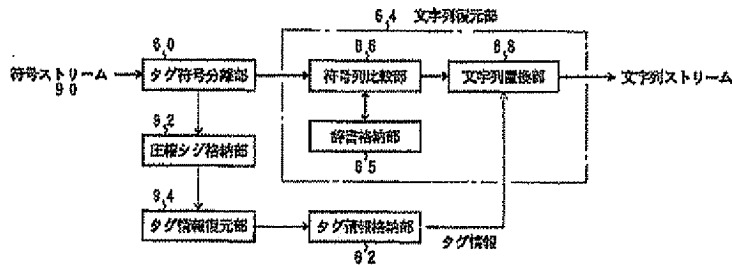


【図18】



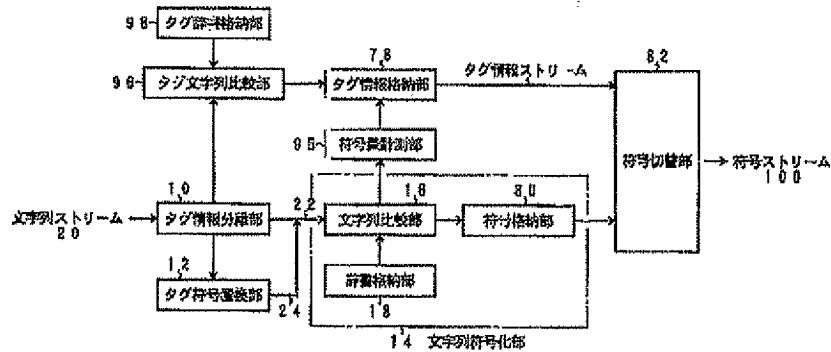
【図19】

図17のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第2実施形態のブロック図

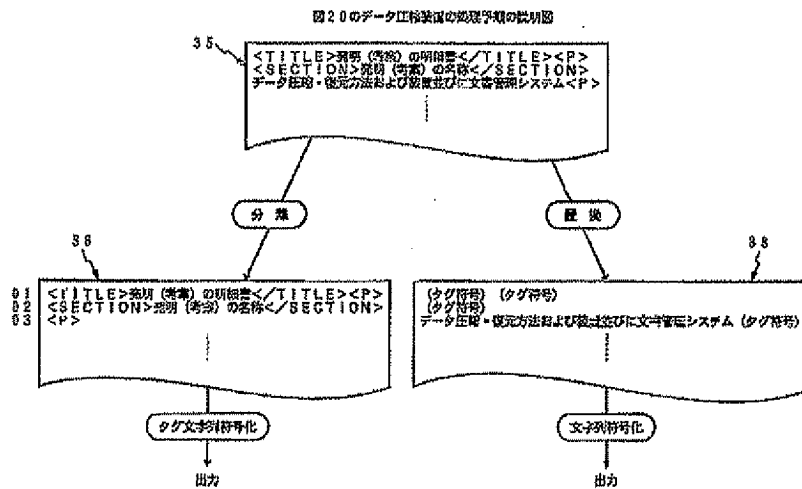


【図20】

本発明のデータ圧縮装置の第4実施形態のブロック図

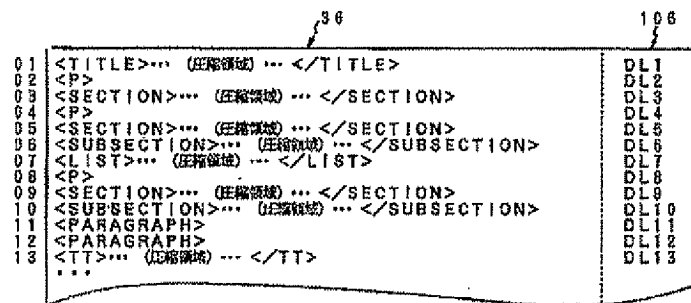


【図21】



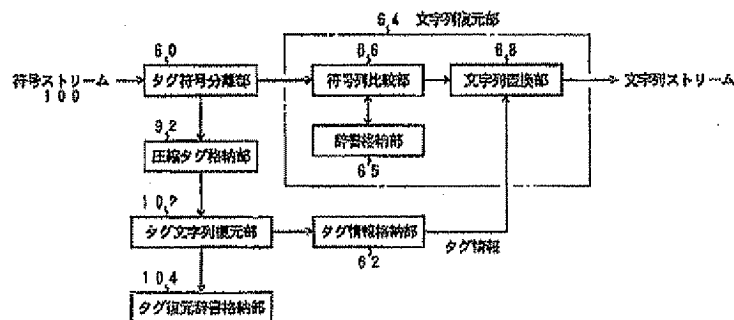
【図22】

図21の符号量をタグに付加した図20のデータ圧縮装置に格納されるタグ情報ファイルとタグ情報ストリームの説明図



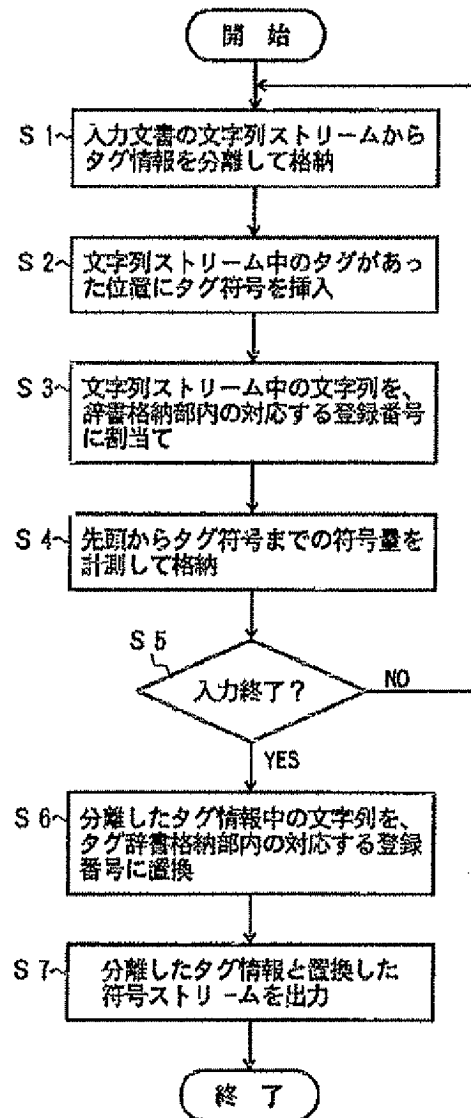
【図24】

図20のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第3実施形態のブロック図



【図23】

図20のデータ圧縮処理のフローチャート



【図27】

日本語文書を例にとったSGML文書ファイルの説明図

```

<TITLE>説明 (表紙) の明細書</TITLE>
<SECTION>説明 (表紙) の名称</SECTION>
データ圧縮・復元方法および圧縮並びに文章管理システム<P>
<SECTION>請求項目の範囲</SECTION>
<SUBSECTION>入力データを符号化するデータ圧縮装置において、</SUBSECTION>
<LIST>
<ITEM>符号化を支持する手段と
<ITEM>符号化を元に入力データを符号化する過程と
</LIST>を有することを特徴としたデータ圧縮方法、<P>
<SECTION>説明の詳細な説明</SECTION>
<SUBSECTION>産業上の利用分野</SUBSECTION>
<PARAGRAPH>近年、文字コード、ベクトル情報、画像など様々な種類のデータがコンピュータ
で扱われるようになっており、扱われるデータ量も急速に増加している。大量のデータを扱うと、
データの中の冗長な部分を省いてデータ量を圧縮することで、記憶容量をへらしたり、速く伝送したり
できるようになる。
<PARAGRAPH>一方、最近では計算機で扱う文章の形式を統一する動きがある。これまで、計算機
あるいはアプリケーションによってばらばらであった文章の形式を、ある計算機でも使用できるようにする
ものである。</P></SGML (Standard Generalized Markup Language) </TT>は1986年にI S

```

【図29】

圧縮していないタグ情報の部分と圧縮した本文部分が混在したSGML文書圧縮データファイルの説明図

```

<TITLE>説明 (表紙) の明細書</TITLE><P>
<SECTION>説明 (表紙) の名称</SECTION>
2f5f6b9d8... (圧縮領域) ...05fc312903
<SECTION>請求項目の範囲</SECTION>
<SUBSECTION>入力データを符号化するデータ圧縮装置において、</SUBSECTION>
6ef268cafd... (圧縮領域) ...358eac0e06fb
<SECTION>説明の詳細な説明</SECTION>
<SUBSECTION>産業上の利用分野</SUBSECTION>
338eac0e06fb... (圧縮領域) ...1f6d38e0c6

```

【手続補正書】

【提出日】平成11年7月14日 (1999. 7. 14)

【手続補正1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項1】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部と、
前記タグ情報分離部でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換部と、
前記タグ符号置換部から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文

字列符号化部と、を有することを特徴とするデータ圧縮装置。

【請求項2】請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮装置。

【請求項3】請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、前記タグ情報分離部で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮装置。

【請求項4】請求項1記載のデータ圧縮装置に於いて、更に、
前記タグ情報分離部で分離されたタグ情報を格納するタグ情報格納部と、
前記文字列符号化部で生成された符号データを格納する符号格納部と、
前記タグ情報格納部に格納されたタグ情報と符号格納部

に格納された符号データを選択して出力する符号切替部と、を設けたことを特徴とするデータ圧縮装置。

【請求項5】請求項1記載のデータ圧縮装置に於いて、前記文字列符号化部は、圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部と、前記タグ符号置換部からの文字列ストリームの中の部分文字列と前記辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部と、を備えたことを特徴とするデータ圧縮装置。

【請求項6】請求項1記載のデータ圧縮装置に於いて、更に、前記タグ情報分離部で分離したタグ情報を圧縮するタグ情報圧縮部を設けたことを特徴とするデータ圧縮装置。

【請求項7】請求項1記載のデータ圧縮装置に於いて、更に、圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、前記タグ情報分離部で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部と、を備えたことを特徴とするデータ圧縮装置。

【請求項8】請求項4記載のデータ圧縮装置に於いて、更に、前記文字列符号化部で生成した符号データの中のタグ位置を検出するタグ位置検出部を設け、前記タグ情報格納部に前記タグ情報分離部で分離したタグ情報と共に前記タグ位置検出部で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮装置。

【請求項9】請求項8記載のデータ圧縮装置に於いて、前記タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納部にタグ情報と共に格納したことを特徴とするデータ圧縮装置。

【請求項10】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置に於いて、前記符号ストリームからタグ情報と符号データとを分離するタグ情報分離部と、前記タグ情報分離部で分離したタグ情報を格納するタグ情報格納部と、前記符号データから文字列及びタグ符号を含む文字列データを復元した後に、前記タグ符号をタグ情報格納部のタグ情報に置き換える文字列復元部と、を備えたことを特徴とするデータ復元装置。

【請求項11】請求項10記載のデータ復元装置に於いて、前記文字列復元部は、

復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部と、前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書格納部の参照で元の文字列を復元する文字列比較部と、

前記文字列比較部により復元したタグ符号を、前記タグ情報格納部のタグ情報に置き換える文字列置換部と、を備えたことを特徴とするデータ復元装置。

【請求項12】請求項10記載のデータ復元装置に於いて、更に、前記タグ情報格納部に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設けたことを特徴とするデータ復元装置。

【請求項13】請求項10記載のデータ復元装置に於いて、更に、

復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、

前記タグ情報分離部により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部と、を備えたことを特徴とするデータ復元装置。

【請求項14】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法に於いて、

前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、

前記タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、

前記タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を有することを特徴とするデータ圧縮方法。

【請求項15】請求項14記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮方法。

【請求項16】請求項14記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、前記タグ情報分離過程で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮方法。

【請求項17】請求項14記載のデータ圧縮方法に於いて、更に、

前記タグ情報分離過程で分離されたタグ情報を格納するタグ情報格納過程と、

前記文字列符号化過程で生成された符号データを格納する符号格納過程と、

前記タグ情報格納過程に格納されたタグ情報と符号格納過程に格納された符号データを選択して出力する符号切

替過程と、を設けたことを特徴とするデータ圧縮方法。

【請求項18】請求項14記載のデータ圧縮方法に於いて、前記文字列符号化過程は、圧縮する際の処理単位となる文字列を登録した辞書を生成する辞書生成過程と、前記タグ符号置換過程で得られた文字列ストリームの中の部分文字列と前記辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。

【請求項19】請求項14記載のデータ圧縮方法に於いて、更に、前記タグ情報分離過程で分離したタグ情報を圧縮するタグ情報圧縮過程を設けたことを特徴とするデータ圧縮方法。

【請求項20】請求項14記載のデータ圧縮方法に於いて、更に、圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を生成するタグ辞書生成過程と、前記タグ情報分離過程で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。

【請求項21】請求項17記載のデータ圧縮方法に於いて、更に、前記文字列符号化過程で生成した符号データのタグ位置を検出するタグ位置検出過程を設け、前記タグ情報分離過程で分離したタグ情報と共に前記タグ位置検出過程で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮方法。

【請求項22】請求項21記載のデータ圧縮方法に於いて、前記タグ位置検出過程は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納過程で分離したタグ情報と共に格納することを特徴とするデータ圧縮方法。

【請求項23】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法に於いて、前記タグ情報と符号データとを分離するタグ情報分離過程と、前記タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、前記符号データから文字列及びタグ符号を含む文字列ストリームを復元した後に、前記タグ符号を前記タグ情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備えたことを特徴とするデータ復元方法。

【請求項24】請求項23記載のデータ復元方法に於いて、

前記文字列復元過程は、復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を生成する辞書生成過程と、前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書の参照で元の文字列を復元する文字列比較過程と、前記文字列比較過程により復元したタグ符号を、前記タグ情報格納過程で分離したタグ情報に置き換える文字列置換過程と、を備えたことを特徴とするデータ復元方法。

【請求項25】請求項23記載のデータ復元方法に於いて、更に、前記タグ情報格納過程で格納されたタグ情報の圧縮データを復元するタグ情報復元過程を設けたことを特徴とするデータ復元方法。

【請求項26】請求項23記載のデータ復元方法に於いて、更に、復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を生成するタグ辞書生成過程と、前記タグ情報分離過程により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書の参照で元のタグ文字列を復元するタグ文字列比較過程と、を備えたことを特徴とするデータ復元方法。

【手続補正2】

【補正対象書類名】明細書

【補正対象項目名】0005

【補正方法】変更

【補正内容】

【0005】このような文書に構造の概念を取り入れた構造化文書の例としては、国際規格のODA (ISO 8613: Open Document Architecture) や、SGML (ISO8879: Standard Generalized Markup Language) の規格による構造化文書がある。またこのような構造化文書を用いた文書処理方法は、例えば特開平5-135054号のものがある。

【手続補正3】

【補正対象書類名】明細書

【補正対象項目名】0013

【補正方法】変更

【補正内容】

【0013】ステップS1で同一の登録文字又は文字列が検索できなかった場合は、ステップS5で元の入力文字又は文字列をそのまま出力する。このような処理をステップS4で入力文字列がなくなるまで繰り返す。

【手続補正4】

【補正対象書類名】明細書

【補正対象項目名】0060

【補正方法】変更

【補正内容】

【0060】続いてステップS3で、タグ配置済みの文

字列ストリーム中の文字列に文字列符号化部14に設けている文字列比較部16で辞書格納部18内の対応する登録番号を符号として割り当て、符号ストリーム26を出力する。このステップS1～S3の処理を、ステップS4で文字列ストリームの入力終了するまで繰り返す。

【手続補正5】

【補正対象書類名】明細書

【補正対象項目名】0062

【補正方法】変更

【補正内容】

【0062】図2の文字列符号化部14に設けた文字列比較部16は、辞書格納部18の参照により、単語を構成する文字列ごとに予め定められた所定の文字列符号を割り当てる符号化を行う。

【手続補正6】

【補正対象書類名】明細書

【補正対象項目名】0064

【補正方法】変更

【補正内容】

【0064】文字列比較部16は日本語文書データの文字列を先頭から順番に入力し、辞書格納部18に予め登録されている単語単位の登録文字列と一致するか否かを検出する。文字列比較部16で入力文字列に一致する登録文字列が検出されると、辞書格納部18の一致検出された登録文字列に対応して予め登録されている文字列符号を読み出して割り当て、この文字列符号を出力する。

【手続補正7】

【補正対象書類名】明細書

【補正対象項目名】0065

【補正方法】変更

【補正内容】

【0065】ここで日本語文書データの文字列を単語単位に文字列符号に変換するための辞書格納部18を説明する。

【手続補正8】

【補正対象書類名】明細書

【補正対象項目名】0069

【補正方法】変更

【補正内容】

【0069】図10は、図2の辞書格納部18の辞書構造の実施形態である。図2の辞書格納部15に格納された辞書は、先頭文字格納部40と従属文字列格納部42の2階層構造を備える。先頭文字格納部40は、日本語

$$K = (N \cdot X - A1) / M$$

但し、X：従属文字列格納部42の位置アドレス

N：一致検出された従属文字列の番号(1, 2, 3, …, N)

A1：従属文字列格納部の開始アドレス

M：従属文字列格納部の格納バイト長

文字「あ、い、う、え、お・・・」の文字コードをインデックスとしており、日本語の文字コードは2バイトデータであることから、文字コード44としては、16進数で「0x0000」から「0xFFFF」の131, 072種類の格納位置が割り当てられる。

【手続補正9】

【補正対象書類名】明細書

【補正対象項目名】0070

【補正方法】変更

【補正内容】

【0070】この文字コード44は、図2の文字列比較部16で読み込んだ先頭文字を使用して、対応する文字コードの位置にアクセスする。文字コード44に続いては先頭アドレス46が格納される。先頭アドレス46は、例えば文字コード44の先頭文字「あ」を例にとると、先頭文字「あ」に続く従属文字列を格納した従属文字列格納部42の先頭アドレス「A1」を指定している。続いて従属文字列の個数48が設けられる。例えば先頭文字「あ」にあつては、従属文字列個数48としてN1=4個が格納されている。

【手続補正10】

【補正対象書類名】明細書

【補正対象項目名】0071

【補正方法】変更

【補正内容】

【0071】従属文字列格納部42は、先頭文字格納部40の先頭文字の文字コード44に対応して格納された先頭アドレス46で先頭位置が指定され、この先頭位置から従属文字列格納部42で指定された個数の格納位置に従属文字列が格納されている。例えば先頭文字「あ」に対応した先頭アドレス46のアドレスA1から従属文字列個数48のN1=4個となる4つの格納位置が、対象とする従属文字列格納領域として指定される。

【手続補正11】

【補正対象書類名】明細書

【補正対象項目名】0074

【補正方法】変更

【補正内容】

【0074】ここで図10の従属文字列格納部42に文字列コード34は、単語個数に基づき1番から136, 486番まで予め17ビットの文字列コードが割り当てられており、図10のように格納した場合の文字列コード(文字列符号)Kと位置アドレスXとの関係は、次式で表すことができる。

(1)

ここで、従属文字列格納部42の格納バイト長Mは、従属文字列の長さ50、従属文字列52、及び文字列コード54の合計長であることから、例えば次式で表すことができる。

$$\begin{aligned}
 \text{格納バイト長} M &= \text{長さ} + \text{文字コード列} + \text{文字列コード} \quad (2) \\
 &= 3\text{bit} + 96\text{bit} + 17\text{bit} \\
 &= 116\text{bit} \\
 &= 15\text{byte}
 \end{aligned}$$

ここでは格納可能な従属文字列を最大6文字とすることで、従属文字列52に96ビットを割り当てた場合を例にとっている。もちろん、従属文字列の文字数は平均で2.8文字であることから、3文字(48ビット)以上とすれば十分な圧縮効果が得られる。この場合には、従属文字列格納部の1つの格納領域の格納バイト長MはM=12バイトになる。

【手続補正12】

【補正対象書類名】明細書

【補正対象項目名】0076

【補正方法】変更

【補正内容】

【0076】図11は図10の辞書構造を持った辞書格納部18による図2の文字列比較部16による符号化処理のフローチャートである。

【手続補正13】

【補正対象書類名】明細書

【補正対象項目名】0079

【補正方法】変更

【補正内容】

【0079】登録した従属文字列と一致すれば、ステップS8に進み、次の文字列コード54を読み出し、一致検出した文字列に文字列比較部16で割り当てて出力し、次のステップS9で先頭文字位置Pを従属文字列の文字数Lだけ移動した位置Pにポインタを更新する。そしてステップS12で非圧縮データの処理が済んでいなければ、再びステップS2に戻り、更新した先頭文字位置Pについて同様な処理を繰り返す。

【手続補正14】

【補正対象書類名】明細書

【補正対象項目名】0082

【補正方法】変更

【補正内容】

【0082】図12は、図2のデータ圧縮装置から出力された符号ストリーム26及びタグ情報ストリーム28で構成される符号ストリームから文字列ストリームを復元するためのデータ復元装置の第1実施形態のブロック図である。

【手続補正15】

【補正対象書類名】明細書

【補正対象項目名】0083

【補正方法】変更

【補正内容】

【0083】このデータ復元装置は、タグ情報分離部60、タグ情報格納部62、文字列復元部64で構成され

る。文字列復元部64は、符号列比較部66、辞書格納部65及び文字列置換部68を備える。

【手続補正16】

【補正対象書類名】明細書

【補正対象項目名】0084

【補正方法】変更

【補正内容】

【0084】タグ情報分離部60は図2のデータ圧縮装置側から送られてきた符号ストリーム56を入力し、タグ情報と符号データとに分離し、タグ情報はタグ情報格納部62に格納し、符号データは符号ストリーム58として文字列復元部64に出力する。

【手続補正17】

【補正対象書類名】明細書

【補正対象項目名】0085

【補正方法】変更

【補正内容】

【0085】文字列復元部64は符号列比較部66で辞書格納部65を用いて符号データから文字列及びタグ符号を復元した後に、文字列置換部68においてタグ符号をタグ情報格納部62に格納しているタグ情報に置き換えて、復元した文字列ストリーム70を出力する。

【手続補正18】

【補正対象書類名】明細書

【補正対象項目名】0086

【補正方法】変更

【補正内容】

【0086】図13は、図12のデータ復元装置の復元処理のフローチャートである。まずステップS1で、タグ情報分離部60が入力文書に対応した符号ストリーム56からタグ情報を分離してタグ情報格納部62に格納する。次にステップS2で、タグ情報が分離された符号ストリーム56の中の符号列を辞書格納部65内の登録番号と比較照合し、一致する登録番号で格納している文字または文字列に変換する。

【手続補正19】

【補正対象書類名】明細書

【補正対象項目名】0088

【補正方法】変更

【補正内容】

【0088】図12の文字列復元部64に設けた符号列比較部66は、辞書格納部65の参照により、図3のデータ圧縮装置で符号化された符号列ストリームから元の文字列を復元する。

【手続補正20】

【補正対象書類名】明細書

【補正対象項目名】0089

【補正方法】変更

【補正内容】

【0089】図14は、図12の文字列辞書格納部65の辞書構造である。この文字列辞書格納部65において

$$\begin{aligned} \text{格納バイト長} M &= \text{先頭文字} + \text{長さ} + \text{文字コード列} & (6) \\ &= 16\text{bit} + 3\text{bit} + 96\text{bit} \\ &= 115\text{ビット} \\ &= 15\text{byte} \end{aligned}$$

から判明しているため、次式から文字列コードKに対応

$$X = M \cdot K + A1$$

但し、K : 文字列コード

A1 : 文字列格納位置の開始アドレス

M : 格納バイト長

このようにして分離した文字列コードKから辞書格納位置を示す位置アドレスXを求めて参照することで、対応する先頭文字及び従属文字列を組み合わせた文字列を復元することができる。

【手続補正21】

【補正対象書類名】明細書

【補正対象項目名】0093

【補正方法】変更

【補正内容】

【0093】タグ情報格納部78にはタグ情報分離部10により文字列ストリーム20から分離されたタグ情報が格納される。これによってタグ情報格納部78には、例えば図6のようなタグ情報ファイル36が格納される。また符号格納部80は文字列符号化部14に設けられており、タグ符号置換部12により分離したタグ情報にタグ情報を挿入したタグ置換済み文字列ストリーム22につき、図11の符号化処理により生成された符号データが格納される。

【手続補正22】

【補正対象書類名】明細書

【補正対象項目名】0097

【補正方法】変更

【補正内容】

【0097】文字列ストリームの入力終了するとステップS5に進み、分離したタグ情報とタグ符号に変換して符号化した符号ストリームを、例えばタグ情報格納部78と符号格納部80から順番に読み出して符号列ストリーム84として出力する。図15のデータ圧縮装置から出力された符号列ストリーム84は、図12に示したデータ復元装置に入力することで文字列ストリームを復元することができる。

【手続補正23】

【補正対象書類名】明細書

【補正対象項目名】0101

は、先頭文字72、従属文字列長さ74及び従属文字列76を、図10の辞書構造に示した従属文字列格納部42の17ビットの文字列コード54の順番に格納している。このため符号列比較部66においては、復元に使用する従属文字列格納部42の格納バイト長Mが

した位置アドレスXを算出することができる。

$$(7)$$

【補正方法】変更

【補正内容】

【0101】タグ情報分離部10、タグ符号置換部12、文字列符号化部14は、図15の第2実施形態と同じである。

【手続補正24】

【補正対象書類名】明細書

【補正対象項目名】0104

【補正方法】変更

【補正内容】

【0104】図19は、図17のデータ圧縮装置から出力された符号ストリーム90から文字列ストリームを復元する本発明のデータ復元装置の第2実施形態である。このデータ復元装置は、図12の第1実施形態に更に圧縮タグ格納部92とタグ情報復元部94を設けている。

【手続補正25】

【補正対象書類名】明細書

【補正対象項目名】0105

【補正方法】変更

【補正内容】

【0105】タグ情報分離部92は、入力する符号ストリーム90に含まれる圧縮タグ情報を分離して圧縮タグ格納部62に格納する。圧縮タグ格納部92に格納された圧縮タグ情報はタグ情報復元部92により復元され、タグ情報格納部62に格納される。タグ情報復元部92はデータ圧縮側のLZ77、LZ78、算術復号化に対応した復元アルゴリズムを実行する。それ以外の構成は図15と同じになる。

【手続補正26】

【補正対象書類名】明細書

【補正対象項目名】0107

【補正方法】変更

【補正内容】

【0107】図20において、タグ情報分離部10、タグ符号置換部12、文字列比較部16、辞書格納部18を備えた文字列符号化部14、タグ情報格納部78及び符号切替部82は、図15の第2実施形態と同じである。

る。これに加えて図20の第4実施形態にあつては、新たにタグ文字列比較部97、タグ辞書格納部96及び符号量計測部98を設けている。

【手続補正27】

【補正対象書類名】明細書

【補正対象項目名】0108

【補正方法】変更

【補正内容】

【0108】タグ文字列比較部97とタグ辞書格納部96は、タグ情報分離部10で分離したタグ情報に含まれる日本語文字列ストリームを文字列符号化部14と同様な符号化アルゴリズムで符号化してタグ情報を圧縮する。このため、タグ情報格納部98の辞書構成は図10と同じであり、先頭文字及び従属文字としてタグ情報に使用する日本語文字列が使用されている。またタグ文字列の符号化処理は図11のフローチャートに従って行う。

【手続補正28】

【補正対象書類名】明細書

【補正対象項目名】0111

【補正方法】変更

【補正内容】

【0111】図21は、図20の第4実施形態における圧縮処理の説明図である。SGML日本語文書ファイル35の内容となる文字列ストリームを入力して、タグ情報の分離によるタグ情報ファイル36の生成及びタグ情報をタグ符号に置換したタグ置換済み日本語文書ファイル38の生成は、図15の第2実施形態と同じである。

【手続補正29】

【補正対象書類名】明細書

【補正対象項目名】0113

【補正方法】変更

【補正内容】

【0113】図22は、タグ情報格納部78に格納されたタグ情報ファイルの具体例であり、図27に示したSGML日本語文書ファイルから分離したタグ情報を例にとっている。このタグ情報ファイル36には、左側のインデックス01～13に対応した各タグに対応して、右側に図21のタグ置換済み日本語ファイル38の文字列データの符号データの先頭からの符号量（バイト量）DL1～DL13が位置指定情報106としてそれぞれ格納されている。

【手続補正30】

【補正対象書類名】明細書

【補正対象項目名】0116

【補正方法】変更

【補正内容】

【0116】このようなステップS1～S4の処理を、ステップS5で文字列ストリームの入力終了するまで繰り返す。文字列ストリーム20の入力が終了すると、ステップS6で、タグ情報格納部78に分離して格納しているタグ情報中の文字列をタグ辞書格納部96内の辞書の対応するブロック番号に変換して符号データとする符号化処理をタグ文字列比較部97で行い、タグ情報格納部78に格納する。その結果、タグ情報格納部78の格納内容は図22の圧縮タグ情報ファイル36のようになる。

【手続補正31】

【補正対象書類名】明細書

【補正対象項目名】0117

【補正方法】変更

【補正内容】

【0117】最後にステップS7で、タグ情報格納部78に分離して符号化した符号量付きのタグ情報と符号格納部80に格納した符号データを符号切替部82より例えば順番に選択出力し、符号ストリーム100として外部に供給する。

【手続補正32】

【補正対象書類名】明細書

【補正対象項目名】0123

【補正方法】変更

【補正内容】

【0123】一方、タグ情報分離部60は、圧縮タグ情報ストリームに続いて送られてくる文書本文の符号ストリームを文字列復元部64に供給し、符号列比較部66で取り出した符号による辞書格納部65の辞書番号の参照で対応する文字または文字列を復元し、文字列置換部68に出力する。

【手続補正33】

【補正対象書類名】明細書

【補正対象項目名】0124

【補正方法】変更

【補正内容】

【0124】文字列置換部68は、復元した文字列の中のタグ符号を認識し、その出現順に従ってタグ情報格納部62に格納している復元済みのタグ情報を格納順に取り出し、タグ符号と置換し、復元した文字列ストリームを出力する。

【手続補正34】

【補正対象書類名】図面

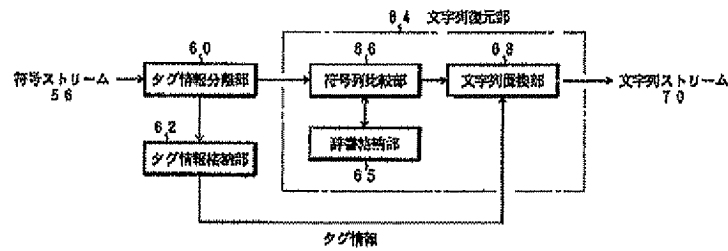
【補正対象項目名】図12

【補正方法】変更

【補正内容】

【図12】

図2のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第1実施形態のブロック図



【手続補正35】

【補正方法】変更

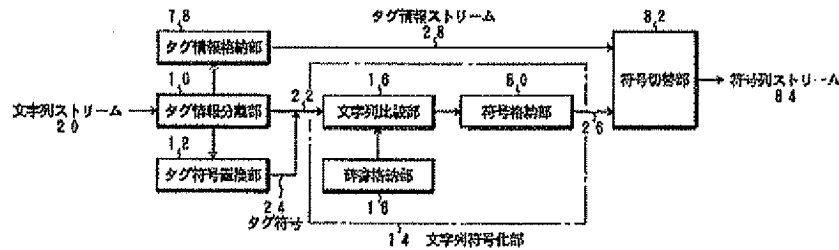
【補正対象書類名】図面

【補正内容】

【補正対象項目名】図15

【図15】

本発明のデータ圧縮装置の第2実施形態のブロック図



【手続補正36】

【補正方法】変更

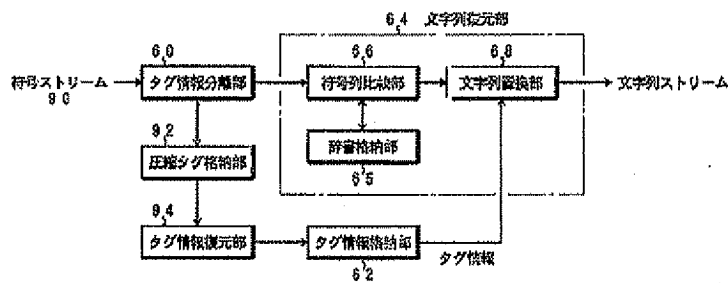
【補正対象書類名】図面

【補正内容】

【補正対象項目名】図19

【図19】

図17のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第2実施形態のブロック図



【手続補正37】

【補正方法】変更

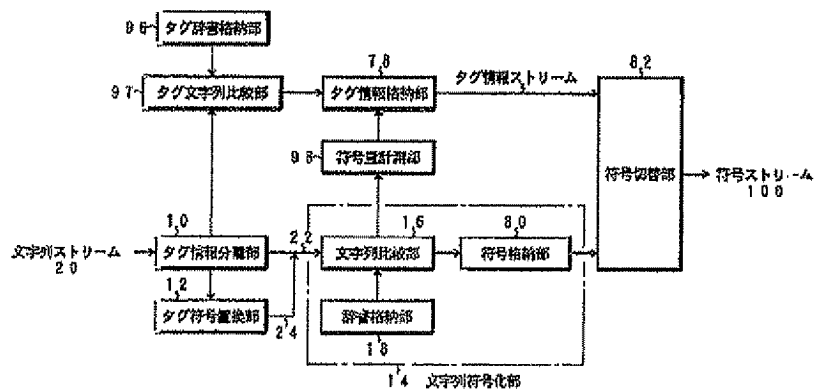
【補正対象書類名】図面

【補正内容】

【補正対象項目名】図20

【図20】

本発明のデータ圧縮装置の第4実施形態のブロック図



フロントページの続き

(72)発明者 佐藤 宣子
神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内